



NATIONAL RESEARCH  
UNIVERSITY

## **Лекция по эконометрике № 4, 3 модуль**

# **Эндогенность, инструментальные переменные**

**Демидова**

**Ольга Анатольевна**

**[https://www.hse.ru/staff/demidova\\_olga](https://www.hse.ru/staff/demidova_olga)**

**E-mail:demidova@hse.ru**

**01.02.2021**

- 1) Линейная регрессия в случае стохастических регрессоров**
- 2) Обобщение теоремы Гаусса-Маркова на случай стохастических регрессоров**
- 3) Проблема эндогенности, несостоятельность оценок МНК**
- 4) Метод инструментальных переменных**
- 5) Двухшаговый МНК**
- 6) Проверка необходимости использования инструментов, тесты Хаусмана и Ву-Хаусмана.**

## Теорема Гаусса-Маркова

Формулировка теоремы при детерминированной матрице  $X$

Модель  $Y = X\beta + \varepsilon$  правильно специфицирована,

$X$  – детерминированная матрица полного ранга,

$$E(\varepsilon) = 0$$

$$\text{Var}(\varepsilon) = \sigma_\varepsilon^2 I_n$$

При выполнении этих предположений оценки МНК являются BLUE.

B – best, L – linear, U – unbiased, E – estimator,

$\hat{\beta}_{OLS}$  – наилучшие линейные несмещенные оценки.

## Ослабление условий теоремы Гаусса-Маркова

Модель  $Y = X\beta + \varepsilon$  правильно специфицирована.

$X$  – случайная матрица полного ранга при любой реализации  $X$ ,

$$E(\varepsilon|X) = 0$$

$$\text{Var}(\varepsilon|X) = \sigma_\varepsilon^2 I_n$$

При выполнении этих предположений оценки МНК являются BLUE.

## Ослабление условий теоремы Гаусса-Маркова

**Более слабое условие: Если**

$$p \lim_{n \rightarrow \infty} \frac{1}{n} X' \varepsilon = 0 \Rightarrow p \lim_{n \rightarrow \infty} \hat{\beta}_n = \beta - \text{состоятельность}$$

**Доказательство.**

$$\begin{aligned} \hat{\beta}_{OLS} &= (X'X)^{-1} X'Y = \beta + (X'X)^{-1} X'\varepsilon, \\ p \lim_{n \rightarrow \infty} \hat{\beta}_{OLS} &= \beta + p \lim_{n \rightarrow \infty} \left[ \left( \frac{X'X}{n} \right)^{-1} \frac{X'\varepsilon}{n} \right] = \\ &= \beta + p \lim_{n \rightarrow \infty} \left[ \left( \frac{X'X}{n} \right)^{-1} \right] p \lim_{n \rightarrow \infty} \left[ \frac{X'\varepsilon}{n} \right] = \beta \end{aligned}$$

## Проблема эндогенности

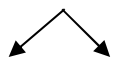
**Если**  $p \lim_{n \rightarrow \infty} \frac{1}{n} X' \varepsilon \neq 0 \Rightarrow p \lim_{n \rightarrow \infty} \hat{\beta}_n \neq \beta$  – несостоятельность

**Доказательство.**

$$\hat{\beta}_{OLS} = (X'X)^{-1} X'Y = \beta + (X'X)^{-1} X'\varepsilon,$$

$$p \lim_{n \rightarrow \infty} \hat{\beta}_{OLS} = \beta + p \lim_{n \rightarrow \infty} \left[ \left( \frac{X'X}{n} \right)^{-1} \frac{X'\varepsilon}{n} \right] =$$

$$= \beta + p \lim_{n \rightarrow \infty} \left[ \left( \frac{X'X}{n} \right)^{-1} \right] p \lim_{n \rightarrow \infty} \left[ \frac{X'\varepsilon}{n} \right] \neq \beta$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$


$$X \uparrow \Rightarrow \varepsilon \uparrow,$$

$$X \uparrow 1 \text{ единицу} \not\Rightarrow Y \uparrow \beta_1 \text{ единиц.}$$

## Причины возникновения эндогенности

- Пропущенные переменные
- Ошибки измерения
- Одновременность
- Самоотбор
- Лаги зависимых переменных (временные ряды)



**Пример:**

$$\text{EARNINGS} = \beta_0 + \beta_1 * \text{EDUCATION} + \dots + \varepsilon$$

**Пропущена переменная «способности», попадающая в  
ошибки, эта переменная коррелирует с переменной  
EDUCATION**

## Ошибки измерения в зависимой переменной

$$Y^* = Y + u,$$

$$Y^* = X\beta + \underbrace{(\varepsilon + u)}_{\varepsilon^*},$$

$\hat{\beta}$  остается *состоятельной*, но  
 $\text{var}(\varepsilon^*) = \text{var}(\varepsilon) + \text{var}(u) \uparrow$

## Ошибки измерения в независимых переменных

$X^* = X + V$ ,  $V$  – матрица *ошибок*.

$$Y = \underbrace{X}_{X^* - V} \beta + \varepsilon,$$

$$Y = X^* \beta + \varepsilon^*, \quad \varepsilon^* = \varepsilon - V\beta,$$

$$\text{cov}(X^*, \varepsilon^*) \neq 0$$

Оценки МНК  $\hat{\beta}$  не являются *состоятельными*.

**Пример: М.Фридмен, критика стандартной функции потребления**  
как линейной зависимости потребления от дохода (общего, а надо от постоянного, общий доход = постоянный + переменный (ошибка измерения)). Склонность к потреблению оказывается завышенной.

## Пример. Из статьи П.Эбес, с.5, Квантиль, 2007

### Системы одновременных уравнений

Обычный (или иерархичный) регрессионный анализ не подходит в случае, когда переменные в правой части модели определяются одновременно с зависимыми переменными. Однако часто бывает трудно избавиться от подобного взаимного влияния. Примером может быть экономический агент, принимающий решения относительно образования или участия на рынке труда (Card, 1999, 2001) или установление цен фирмами в условиях конкуренции. Некоторые исследования рассматривают одновременность цены и величины спроса на рынках с дифференцированным продуктом при данной структуре конкуренции. Ценовая политика фирм, обусловленная, например, ненаблюдаемыми характеристиками товара, такими как наличие скидок и общенациональной рекламы, расположение точек продаж и другими параметрами розничной торговли, или же реакция со стороны конкурентов, ведет к эндогенности. Работа Berry (1994) по борьбе с эндогенностью цен в агрегированных моделях с использованием инструментальных переменных широко используется и адаптируется. Например, Nevo (2001) оценивает структурную модель спроса и предложения для отрасли готовых к употреблению зерновых завтраков; Berry, Levinsohn & Pakes (1995) и Sudhir (2001) разрабатывают модель рыночного равновесия с конкурентным ценообразованием на рынке автомобилей для исследования ценообразования на автомобили и уровня конкуренции. Свежий обзор структурного моделирования в маркетинге содержится в Chintagunta, Erdem, Rossi & Wedel (2006).

Простая модель спроса и предложения для продукта или товара выглядит следующим образом:

$$\begin{aligned}y_t^d &= (x_t^d)' \beta^d + \gamma^d p_t + \epsilon_t^d, \\y_t^s &= (x_t^s)' \beta^s + \gamma^s p_t + \epsilon_t^s,\end{aligned}$$

где компоненты вектора  $x_t^d$  – это факторы, влияющие на спрос или поведение потребителей, а компоненты  $x_t^s$  влияют только на поведения производителей. Цена  $p_t$  определяется из равенства  $y_t^d = y_t^s$ . Когда оценивается уравнение спроса  $y_t^d = (x_t^d)' \beta + \gamma^d p_t + \epsilon_t^d$ , нельзя предполагать, что  $E[\epsilon_t^d | p_t] = 0$ , так как цена и величина спроса определяются одновременно, то есть ненаблюдаемые положительные шоки спроса или действия конкурентов сдвигают кривую спроса вверх, что (при прочих равных) означает более высокую равновесную цену. В этом случае МНК нельзя использовать для получения оценок параметров уравнения спроса.

## Пример. Из статьи П.Эбес. с.4. Квантиль. 2007

### Самоотбор

Проблема самоотбора возникает, когда индивиды выбирают себе определенное состояние, например, быть или нет членом профсоюза (Vella & Verbeek, 1998), лечиться или нет (Angrist, Imbens & Rubin, 1996), на основании экономических или других, обычно неизвестных, причин. Например, Angrist (1990) рассматривает влияние статуса ветерана войны во Вьетнаме на доход граждан, чтобы понять, следует ли правительству США давать им компенсацию за возможную потерю личного дохода, вызванную службой в армии. Однако доходы не просто сравнить учитывая лишь статус ветерана, потому что индивиды с меньшими возможностями «на гражданке» скорее поступят на военную службу, и такие индивиды зарабатывали бы меньше независимо от службы в армии.

Hamilton & Nickerson (2003) дают обзор эндогенного принятия решений в стратегическом менеджменте, когда менеджеры осуществляют организационный выбор из нескольких конкурирующих стратегий не случайно, а на основании ожиданий и опыта. Аналогично данные, собранные в интернете, могут страдать от проблемы самоотбора. Определенного рода индивиды чаще бывают в сети и, следовательно, чаще принимают участие в онлайн-опросах, заходят на вебсайты или делают покупки в интернет-магазинах. Если эти ненаблюдаемые индивидуальные характеристики влияют на поведение в сети, предпочтения или восприятие, то часть влияния этих скрытых характеристик неправильно приписывается использованию интернета. Можно предположить, что эти индивиды вели бы себя иначе независимо от частоты посещения интернета. Эти явления важны, например, при исследовании решений о количестве приобретаемого товара в интернет-магазинах по сравнению с обычными («оффлайн-овыми») магазинами или о покупке товаров определенного бренда в зависимости от категории товара и характеристик магазина.

Проиллюстрируем простую модель с самоотбором:

$$\begin{aligned} y_i &= x_i'(\beta + \delta) + \epsilon_i & \text{если } i \in I, \\ &= x_i'\beta + \epsilon_i & \text{если } i \in II, \end{aligned}$$

где I и II обозначают определенные состояния (например, интернет-пользователь или нет). Более компактная запись:

$$y_i = x_i'\beta + d_i x_i'\delta + \epsilon_i,$$

где  $d_i = 1$ , если  $i \in I$ , и  $d_i = 0$  в противном случае. Из этой записи видно, что  $d_i$  является фиктивной переменной, и стандартное оценивание не проходит, если  $E[\epsilon_i | d_i] \neq 0$ . Это предположение нарушается в приведенных выше примерах. Более детально о проблеме самоотбора можно узнать, например, из Vella (1998).



## Что делать?

- 1) Метод инструментальных переменных (IV),
- 2) Двухшаговый МНК (2SLS),
- 3) Обобщенный метод моментов (GMM).

Парная регрессия:  $Y = \beta_0 + \beta_1 X + \varepsilon$ ,

Переменные  $Z_1, \dots, Z_L$  называются

инструментальными переменными (или  
инструментами) для  $X$ , если

1)  $|\text{cor}(X, Z_i)| > 0$ ,  $i = 1, \dots, L$  (релевантность).

Если  $|\text{cor}(X, Z_i)|$  невелико, то это слабый инструмент,

Если  $|\text{cor}(X, Z_i)| \approx 1$ , то это сильный инструмент

2)  $\text{cor}(Z_i, \varepsilon) = 0$  (валидность)

Более слабое условие:  $p \lim_{n \rightarrow \infty} Z_i' \varepsilon = 0, \quad i = 1, \dots, L$



## Пример 1 инструментальных переменных

*Example*

$$\text{Score} = \beta_1 + \beta_2 CA + \varepsilon,$$

*CA (class attendance) is endogenous variable*

*Instruments:  $Z_1$  – distance to the University campus,*

*$Z_2$  – transport expenditures*

## Пример 2 инструментальных переменных

$$\text{EARNINGS} = \beta_0 + \beta_1 * \text{EDUCATION} + \dots + \varepsilon$$

Пропущена переменная «способности», попадающая в ошибки, эта переменная коррелирует с переменной EDUCATION,

Инструменты для переменной EDUCATION:

SM – длительность обучения мамы индивида,

SF – длительность обучения папы индивида.

## Пример 3 инструментальных переменных

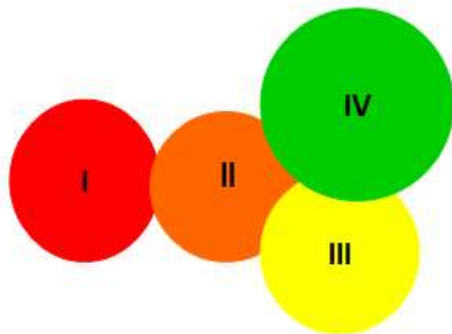
Пространственно-эконометрическая модель:

$$Y = \rho WY + X\beta + \varepsilon,$$

$W$  – взвешивающая матрица,

$WY$  – пространственный лаг зависимой переменной (эндогенная переменная).

Пример взвешивающей матрицы:



$$W_b = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

## Пример 3 инструментальных переменных

Пример пространственного лага:

$$WY = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} = \begin{pmatrix} Y_2 \\ 1/3(Y_1 + Y_3 + Y_4) \\ 1/2(Y_2 + Y_4) \\ 1/2(Y_2 + Y_3) \end{pmatrix}$$

## Пример 3 инструментальных переменных

Пространственно-эконометрическая модель:

$$Y = \rho WY + X\beta + \varepsilon,$$

$WY$  – пространственный лаг зависимой переменной (эндогенная переменная).

Инструменты для  $WY$ :  $X$ ,  $WX$ ,  $W^2X$ .

Обоснование:

$$(I - \rho W)Y = X\beta + \varepsilon,$$

$$Y = (I - \rho W)^{-1} (X\beta + \varepsilon),$$

$$Y = (I + \rho W + \rho^2 W^2 + \dots) (X\beta + \varepsilon)$$

## Оценка инструментальных переменных для коэффициента наклона в парной регрессии

Парная регрессия:  $Y = \beta_0 + \beta_1 X + \varepsilon$ ,

$Z$  - инструмент для  $X$ .

Определение.  $\hat{\beta}_{IV} = \frac{c\hat{v}(Z, Y)}{c\hat{v}(Z, X)}$

Эта оценка состоятельна:

$$\hat{\beta}_{IV} = \frac{c\hat{v}(Z, Y)}{c\hat{v}(Z, X)} = \frac{c\hat{v}(Z, \beta_0 + \beta_1 X + \varepsilon)}{c\hat{v}(Z, X)} = 0 + \beta_1 + \frac{c\hat{v}(Z, \varepsilon)}{c\hat{v}(Z, X)}$$

$$p \lim_{n \rightarrow \infty} \hat{\beta}_{IV} = \beta_1 + p \lim_{n \rightarrow \infty} \frac{c\hat{v}(Z, \varepsilon)}{c\hat{v}(Z, X)} = \beta_1 + \frac{p \lim_{n \rightarrow \infty} \frac{1}{n} Z' \varepsilon}{p \lim_n \frac{1}{n} Z' X} = \beta_1$$

## Оценка инструментальных переменных для множественной регрессии

**Множественная регрессия:  $Y = X\beta + \varepsilon$ ,**

**$Z_1, \dots, Z_m$  - инструменты для  $X_1, \dots, X_k$ .**

**Оценка инструментальных переменных (IV), если  $m=k$ :**

$$\hat{\beta}_{IV} = (Z'X)^{-1} Z'Y$$

**Эта оценка состоятельна:**

$$\begin{aligned}\hat{\beta}_{IV} &= (Z'X)^{-1} Z'Y = (Z'X)^{-1} Z'(X\beta + \varepsilon) = \\ &= \beta + (n^{-1}Z'X)^{-1} n^{-1}Z'\varepsilon\end{aligned}$$

$$p \lim \frac{1}{n} Z'\varepsilon = 0 \qquad p \lim \hat{\beta}_{IV} = \beta$$

Множественная регрессия:  $Y = X\beta + \varepsilon$ ,  
 $Z_1, \dots, Z_m$  - инструменты для  $X_1, \dots, X_k$ .

Оценка 2-х шагового МНК (two stage least square)  
2SLS estimate ( $m > k$ ).

**1<sup>ый</sup> шаг:** Проекция  $X_1, \dots, X_k$  в подпространство  $Z_1, \dots, Z_m$

$$\hat{X}_j = Z(Z'Z)^{-1}Z'X_j$$

**2<sup>ой</sup> шаг:**

$$Y = \beta_0 + \beta_1\hat{X}_1 + \dots + \beta_k\hat{X}_k + \varepsilon$$

$$\hat{\beta} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y$$



**Задание (самостоятельное):**

**Показать, что если  $m = k$ , оценка двухшагового МНК  
совпадает с оценкой инструментальных  
переменных.**

## Надо ли применять оценки ИП?

$$Y = X\beta + \varepsilon$$

$$\hat{\beta}_{OLS} = (X'X)^{-1} X'Y,$$

$$\text{var}[\hat{\beta}_{OLS}] = (X'X)^{-1} X' \text{var}(Y) X (X'X)^{-1} = \sigma_{\varepsilon}^2 (X'X)^{-1},$$

$$\hat{\beta}_{IV} = (Z'X)^{-1} Z'Y,$$

$$\text{var}[\hat{\beta}_{IV}] = (Z'X)^{-1} Z' \text{var}(Y) Z (Z'X)^{-1}$$

**Если на самом деле эндогенности нет, то оценки IV будут неэффективными.**

**Как понять, надо ли использовать оценки IV, 2SLS?**

## Тесты Хаусмана на проверку эндогенности

$$Y = X_1 \beta_1 + X_2 \beta_2 + \varepsilon,$$

$(n \times k_1)$                        $(n \times k_2)$

$X_2$  – матрица экзогенных переменных,

$X_1$  – матрица возможно эндогенных переменных.

$Z_1$  - матрица инструментов для  $X_1$   
размера  $n \times m$ ,  $m \geq k_1$ .

$Z = [Z_1, X_2]$  – матрица всех экзогенных переменных.

## Тест Ву-Хаусмана (Wu – Hausman)

$H_0$  :  $X_1$  и  $\varepsilon$  не коррелируют,  
 $X_1$  - матрица экзогенных переменных

$H_1$  :  $X_1$  и  $\varepsilon$  коррелируют,  
 $X_1$  - матрица эндогенных переменных

*Тест Ву – Хаусмана*

*1 шаг: Регрессия всех переменных из  $X_1$  на  $Z$ :*

$$X_{1j} = Z\gamma_j + v_j, j = 1, \dots, k_1 \Rightarrow \hat{v}_j$$

## Тест Ву-Хаусмана (Wu – Hausman)

*2 шаг: Регрессия*

$$Y = X_1\beta_1 + X_2\beta_2 + \gamma_1\hat{v}_1 + \dots + \gamma_{k_1}\hat{v}_{k_1} + \varepsilon,$$

$$H_0 \Leftrightarrow \gamma_1 = \dots = \gamma_{k_1} = 0$$

$$H_1 : \gamma_1^2 + \dots + \gamma_{k_1}^2 > 0$$

Статистика Вальда  $\chi^2(k_1)$   
или F - статистика

## Тест Хаусмана

$H_0 : X_1$  и  $\varepsilon$  не коррелируют  $\Rightarrow$   
 $\hat{\beta}_{OLS}$  эффективная и состоятельная,  
 $\hat{\beta}_{IV}$  состоятельная,  
 $\Rightarrow$  разница между оценками мала.

$H_1 : X_1$  и  $\varepsilon$  коррелируют,  
 $\hat{\beta}_{OLS}$  несостоятельная,  $\hat{\beta}_{IV}$  состоятельная  
 $\Rightarrow$  разница между оценками велика.

### Равносильный тест Хаусмана

$$q = \hat{\beta}_{IV} - \hat{\beta}_{OLS},$$

Тестовая статистика :

$$q' \left( \text{var}(\hat{\beta}_{IV} - \hat{\beta}_{OLS}) \right)^{-1} q \stackrel{as}{\sim}$$

$$q' \left( \text{var}(\hat{\beta}_{IV}) - \text{var}(\hat{\beta}_{OLS}) \right)^{-1} q \sim \chi^2(k_1)$$



NATIONAL RESEARCH  
UNIVERSITY

# Thank you for your attention!

20, Myasnitskaya str., Moscow, Russia, 101000  
Tel.: +7 (495) 628-8829, Fax: +7 (495) 628-7931  
[www.hse.ru](http://www.hse.ru)